DOCUMENT RESUME

ED 400 316 TM 025 709

AUTHOR Perlman, Carole; And Others
TITLE The Effect of Extended Time Limits on Learning Disabled Students' Scores on Standardized Reading Tests.
PUB DATE 9 Apr 96
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Achievement Tests; Elementary Education; *Elementary School Students; Grade 4; Grade 8; *Learning Disabilities; *Reading Tests; *Scores; *Standardized Tests; Student Attitudes; Student Motivation; Test Format; Test Length; Test Reliability; Test Results; *Timed Tests
IDENTIFIERS Iowa Tests of Basic Skills

ABSTRACT

Eighty-five fourth- and eighth-grade learning disabled students whose individualized education plans specified untimed achievement testing were tested with the Reading Comprehension subtest of the Iowa Tests of Basic Skills, either according to the publisher's 40-minute time limit or with an extended time limit of 2 hours, 30 minutes. Results were compared with those from a parallel form of the test administered 8 months earlier. Measures included grade-equivalent and normal curve equivalent scores, completion rates, amount of time needed to complete the test, and KR-20 reliabilities based on students in each cell. All but one of the fourth graders in the extended time group finished the test in the publisher's time limit. Twenty-two of the 26 eighth graders in the unlimited time group took longer than the publisher's limit. Despite some methodological problems centering around student motivation and attitudes toward testing, the study suggested that the test may be more reliable when administered without time limits, and students' merely knowing that unlimited time is available may yield higher scores, even if no additional time is actually used. (Contains 5 tables and 10 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
CAROLE PERLMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 400 316

The Effect of Extended Time Limits on Learning Disabled Students' Scores on Standardized Reading Tests

by

Carole Perlman
Jeanne Borger
Carla Collins
Janet Elenbogen
Judith Wood

Chicago Public Schools

Paper presented at the Annual Meeting of the
National Council on Measurement in Education, New York, April 9, 1996

TM025709

BEST COPY AVAILABLE

ABSTRACT

The Effect of Extended Time Limits on Learning Disabled Students' Scores on Standardized Reading Tests

Carole L. Perlman, Jeanne Borger, Carla B. Collins, Janet C. Elenbogen, and Judith Wood
Chicago Public Schools

Eighty-five fourth- and eighth-grade learning disabled students whose IEPs specified untimed achievement testing were tested tested with the Reading Comprehension subtest of the *Iowa Tests of Basic Skills* either according to the publisher's 40-minute time limit or with an extended time limit of 2 hours, 30 minutes. Results were compared with those from a parallel form of the test administered eight months earlier. Measures include grade-equivalent and normal curve equivalent scores, completion rates, amount of time needed to complete the test and KR-20 reliabilities based on students in each cell. Despite some methodological problems, the study suggests that that (1) the test may be more reliable when administered without time limits and (2) merely knowing that unlimited time is available may yield higher scores, even if no additional time is actually used.

The Effect of Extended Time Limits on Learning Disabled Students' Scores on Standardized Reading Tests

Background

The Americans with Disabilities Act requires that reasonable and appropriate accommodations be provided when students with disabilities are assessed. One commonly provided accommodation is to allow the student extra time to complete the test. The following excerpt from the 1985 edition of the *Standards for Educational and Psychological Testing* speaks to that issue:

> Although modifications in the time allowed for tests are considered among the appropriate test options, there are few data available to support any conclusions about the effects of modifications in time...on the test results. Furthermore, little is known about how much time people with various handicapping conditions actually need because records of the time actually used are rare and empirical studies to set time limits are even more rare. (p. 77)

Ysseldyke (1995) maintains that little has changed since then and that not much is known about the effects of testing accommodations on the results of those assessments and the implications of modifications on how the results should be interpreted. Geisinger (1994) also points out that the meaning of scores from nonstandard test administrations needs further study, particularly with regard to validity and robustness. What little research has been done has focused on college admissions tests (e.g., Willingham, *et al.*, 1988), rather than on elementary and secondary standardized achievement tests. This leaves states and school systems with little guidance on how to deal with assessments that were administered with accommodations, except to attempt to use criterion-referenced, rather than norm-referenced interpretations of the results (McLoughlin & Lewis, 1981).

The National Center on Educational Outcomes (1994) has noted that students with disabilities are too often excluded from local and state assessment programs. This effectively eliminates such students from being considered when instructional programs and reform efforts are evaluated and when accountability systems are established. Exclusion may come about as a result of a student not being assessed at all or by administering the assessment with accommodations that would prevent the results from being aggregated with scores of students who were assessed under standard conditions.

Learning disabled students comprise nearly half of all students receiving federally funded special education services and, in 1991-92, they represented 5.31% of the total K-12 enrollment (National Center for Education Statistics, 1994, p. 65). One frequently recommended accommodation for learning disabled students is extended time. For example, of the eight states studied by Bell (1993), seven specifically suggested extended time as an accommodation. All of the states leave the decisions about accommodations to local school system personnel; in most instances the student's individualized education program (IEP) is to specify which accommodations are to be used on local and state assessments. However, anecdotal evidence suggests that those decisions are often not made in a systematic way.

Objectives

The objectives of this study were to answer the following questions:

1. Do learning disabled elementary school students score higher on standardized reading tests when the publisher's time limits are exceeded?
2. If so, what is the magnitude of that difference and does it remain constant across grade levels?
3. How long does it take learning disabled students to complete the test?
4. How is the reliability of the test affected by an extended-time administration?

Instruments/Method

A 2x2 factorial design was used (grade 4 vs. grade 8 and regular timing vs. extended time). The subjects were learning disabled students enrolled in a large, ethnically diverse, urban school system. All students in the study had IEPs that recommend that they be given extended time to complete standardized reading tests. Students were enrolled at a total of 19 schools at the school system's benchmark grades, 4 ($N = 28$) and 8 ($N = 57$). With parental permission, they were tested in early January with the Reading Comprehension subtest of the *Iowa Tests of Basic Skills* (ITBS) *Form H* (Hieronyous, *et al.*, 1986), using the same level, but a different form of the test they took the previous May as part of the systemwide testing program (see Table 1); the spring test was the ITBS Form K; both forms used 1988 norms. The only fourth graders who could participate in the study were those who took the ITBS at grade level (Level 9) the previous spring; students who were tested below grade level would have taken a test that would be untimed for *all* students, disabled or not. Just under half of all eighth graders in the study were tested at grade level, though the percentages varied between the timed and untimed groups (29% vs. 58%, respectively). Records were kept of the actual amount of time that each student in the extended-time groups needed to complete the test (all students in the timed groups were given the 40 minutes specified in the publisher's instructions). Tests were given by central office staff, all of whom had experience in giving standardized tests. This was done partly to ensure that the testing was done according to the publisher's guidelines (except for the untimed groups) and partly to avoid creating extra work for the school staff.

Because of unanticipated difficulties in obtaining subjects, the design of the study differed in two important ways from what had been planned. First, although we had planned to test 70 students per grade, we were unable to recruit that many subjects. It was also our intention to randomly assign half the students to take the test using the publisher's 40-minute time limit, while the other half was given up to 2 hours, 30 minutes (this essentially amounted to giving the test without any time limits at all). This design was chosen over one in which students took two forms of the test, one with time limits, one without, in order to minimize disruption of the students' regular schoolwork. However, it did not prove feasible to conduct testing at a given school under both timed and untimed conditions, so entire schools were assigned to test either with normal time

limits or extended ones. This had the effect of confounding treatment with schools' individual testing policies regarding special education students.

The gender distributions for timed vs. untimed groups were similar. The percentage of students tested at grade level was identical for the two fourth grade groups, but differed somewhat between the timed and untimed groups at grade 8. The fourth graders performed relatively better on the spring pretest than did the eighth graders. (Pretest scores were not readily available for examination until *after* the treatment groups were established and posttested.) Fourth graders in the timed condition slightly outperformed untimed students on the pretest, while the two groups of eighth graders scored equally well on the pretest.

Results

Table 1 gives descriptive statistics for the population tested, for testing time, pretest scores, and posttest scores. Fourth graders in the extended time condition took an average of 34 minutes to complete the (normally 40-minute) test; all but one of the fourth graders in the extended time group finished the test within the publisher's time limit. Eighth graders, took substantially longer, an average of 55 minutes; 22 of the 26 students in the untimed group took more than 40 minutes to finish.

An ANCOVA was performed on posttest grade-equivalent scores, with pretest grade-equivalent as the covariate. The correlation between pretest and posttest scores was 0.79. Results of the ANCOVA are presented in Table 2. These results indicate that the students did get higher scores when taking the test with extended time $F(1,80) = 10.55$, $p < 0.01$. Not surprisingly, grade was also significant, $F(1,80) = 15.53$. The grade X treatment interaction was not significant, $F(1,80) = 0.21$, $p = 0.65$. Adjusted means are presented in Table 3.

Coefficient α reliabilities of the pretest and posttest were calculated separately for the students in each cell of the design for the modal test level at each grade (level 9 for grade 4 and level 13 for grade 8). These reliabilities are presented in Tables 4 and 5.

Discussion

Many things about this study are troubling and make it difficult to draw any meaningful conclusions. Some of the problems with the design and possible lack of comparability between the fourth and eighth grade groups and the timed and untimed groups were mentioned in the previous two sections. Although no mention was made to the students that timing was the relevant variable, many did see the letter requesting parental permission that said that the students might take a timed test. It also appeared that some of the students were not very motivated when they took the posttest and a few seemed to respond quite negatively, perhaps because they were missing a favorite activity (such as gym), because they felt singled out, or because they found testing aversive. There was also an artificial quality to the testing situation, since testing wasn't done by school staff and because only learning disabled students were tested.

The issue of test reliability (or the lack thereof) is cause for concern. We have no good explanation for the apparent total unreliability of the pretest measure, especially at eighth grade. (If there is an error in programming, four people have been unable to detect what it might be. That the reliability for the entire population of students tested the previous spring was .88 suggests a problem with reliability within the LD sample, rather than a programming problem.) In any case, if the pretest really is that unreliable, it makes most of the results of this study very hard to interpret. On the posttest, the tests appeared to be more reliable when administered without time limits. This would seem to argue in favor of extended time limits.

It became clear during the course of the study that although there is a systemwide assessment policy for special education students (students should participate in the testing program to the maximum extent possible, tests should be given without major accommodations whenever possible, accommodations should mirror procedures used in regular classroom instruction, and testing decisions should be made on a case-by-case basis), that policy is applied very differently in different schools. For example, some schools may give all of their learning disabled students extra testing time, while other schools offer extended time rarely, if ever. Some schools seem to be more likely than others to test special education students at grade level. And while some take pains to base their testing decisions on each individual student's needs, others seem to make blanket decisions based on the type of disability or simply on the fact that there is a disability. Schools might be understandably reluctant to talk about how they make their testing decisions, especially when those decisions run counter to district-wide policy or in those instances where the courts are heavily involved in monitoring special education practices, as they are in our district. Nonetheless, school level policy decisions are obviously important and deserve study.

Clearly, some testing accommodations are reasonable and appropriate, while others, resting on a shakier foundation, may not be either appropriate or necessary. Since extended time is a serious enough modification that it would normally result in a student's scores being eliminated from school summary statistics, there may be some incentive for the school to err on the side of excluding students from those statistics, especially when the assessment is a high stakes assessment for the teachers or school administrators, e.g., when test results are used to determine merit pay, whether an administrator's contract is renewed, or whether a school maintains its state accreditation. Altruism, too, has a role in this process; some teachers may be reluctant to expose their students to situations that the teacher feels may be stressful, such as taking a timed test. Whatever the cause, the elimination of these students from school summary data is hardly desirable from a policy perspective, since it provides a misleading picture of the achievement level of the school population and because it sends the message that schools will not be held accountable for the achievement of all their students.

Although the results of this study are hardly conclusive, they do suggest that not all students actually use the extra time they are given and possibly that older students are more likely to need extra time than younger ones. It may be the eighth graders have relatively more serious disabilities than the fourth graders. In our school district, the average learning disabled student is nine years old at the time of diagnosis (Elenbogen & Fredrick, 1994) and the average length of time a learning disabled student remains in special education is about 4½ years. Perhaps the students still remaining in special education in eighth grade are those whose disabilities are

particularly severe or resistant to treatment. That the fourth graders in the untimed conditions tended to score higher than those who took the test with the conventional time limit, *even though they both used about the same amount of time,* suggests that the relevant variable may not be the actual amount of time available. The better performance may be instead the result of reduced stress and more positive expectations resulting from the students' perception that they would have all the time they needed. Yet, regardless of the reason, the students in the untimed conditions did tend to perform better.

In some respects this study brought up more questions than it answered. Other, better designed studies need to be conducted, examining a variety of tests, subjects and grades. Ideally, the experimental conditions should more closely approximate those of the regular district-wide or state-wide testing program or even be part of it, IEPs and parental consent permitting. For more than a decade, our district has conducted retesting of randomly selected classrooms with an alternate form of the *Iowa Tests of Basic Skills*; it's been our experience that scores generally are lower on the retest, even in the absence of any inappropriate testing practices or cheating. The most likely cause of the lower scores is the drop from the high level of motivation characteristically present during the regular testing period to one that often could most charitably be described as more relaxed. We've found it very difficult to motivate students during a retest, even if the students' own teacher does the testing and even if the students are told that their scores "count." It is especially important that any research on testing conditions be part of the "real" testing whenever possible.

If the results of these studies indicate that administering standardized tests with conventional time limits does not make a substantial difference in students' scores, perhaps more students will be tested within time limits and therefore, eligible to be counted in school averages. In any case, the results of the studies should provide educators some needed guidance as they craft the assessment sections of students' individualized educational programs.

References

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bell, S. (1993). *Strategies for Exempting or Accommodating Special Populations in Statewide Assessment Programs*. Glen Ellyn, IL: Institute for Educational Research.

Elenbogen, J. & Fredrick, W. (1994) *Age of Identification and Length of Special Education Services*. Chicago: Chicago Public Schools Department of Research, Evaluation & Planning.

Geisinger, K.F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, *7*(2), 121-140.

Hieronymus, A.N., Hoover, H.D., & Lindquist, E.F. (1986). *Iowa Tests of Basic Skills Form H*. Chicago: Riverside Publishing Company.

McLoughlin, J.A. & Lewis, R.B. (1981). Assessing special students. Columbus, OH: Merrill.

National Center for Education Statistics. (1994). *Digest of Education Statistics 1994*. Washington: U.S. Department of Education Office of Educational Research and Improvement.

National Center on Educational Outcomes (1994). *NCEO Policy Directions*, *1* (1). Minneapolis: Author.

Willingham, W.W., Ragosta, M., Bennett, R.E., Braun, H., Rock, D.A., & Powers, D.E. (Eds.). (1988). *Testing Handicapped People*. Needham Heights, Massachusetts: Allyn-Bacon.

Ysseldyke, J.E. (1995, June). Progress and Programs for including students with disabilities in national and state assessments. Presentation to the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix.

Table 1
Descriptive Statistics

Variable		Grade 4		Grade 8	
		Timed (*N* = 16)	Untimed (*N*=12)	Timed (*N*=31)	Untimed (*N*=26)
Gender	male	12 (75%)	11 (92%)	21 (68%)	19 (73%)
	female	4 (25%)	1 (8%)	10 (32%)	7 (27%)
Tested at grade level?	yes	16 (100%)	12 (100%)	9 (29%)	15 (58%)
	no	0 (0%)	0 (0%)	22 (71%)	11 (42%)
Time students took to complete the test (for untimed groups; students in timed group were all given 40 min.)	*M*	-	34.00	-	54.96
	SD	-	6.31	-	16.95
	SE	-	1.82	-	3.32
	median	-	33	-	54
	range	-	22-45	-	27-97
Pretest Grade-Equivalent Score	*M*	2.47	2.22	4.68	4.67
	SD	0.77	0.75	1.21	1.28
	SE	0.19	0.22	0.22	0.25
	median	2.50	2.20	4.70	4.85
Posttest Grade-Equivalent Score	*M*	1.87	2.32	4.36	5.16
	SD	0.93	0.87	1.22	1.31
	SE	0.23	0.25	0.22	0.26
	median	1.70	1.9	4.10	5.25
Pretest Percentile Rank	median	15	9.5	6.0	8.5
	range	1-53	1-53	1-36	1-32
Posttest Percentile Rank	median	1	2	2	9.5
	range	1-45	1-39	1-17	1-48
Pretest Normal Curve Equivalent Score	*M*	25.87	22.08	17.35	19.38
	SD	15.23	14.07	12.69	11.38
	SE	3.93	4.06	2.28	2.23
	median	28.0	22.0	17.0	21.0
Posttest Normal Curve Equivalent Score	*M*	9.00	16.08	10.71	21.15
	SD	13.58	15.70	10.95	15.37
	SE	3.40	4.53	1.97	3.02
	median	1.00	7.00	7.0	22.0

BEST COPY AVAILABLE

Table 2
ANCOVA for Grade-Equivalent Posttest Reading Scores

Variable	*df*	Sum of squares	Mean square	*F*-value	*p*-value
Pretest	1	39.25	39.25	45.28	<.001
Timed/Untimed	1	9.15	9.15	10.55	<.01
Grade	1	13.47	13.47	15.53	<.001
Interaction	1	0.18	0.18	0.21	.65
Residual	80	69.36	0.87		

Table 3
Adjusted Means

	Timing	Grade	*N*	Adjusted mean
Overall	-	-	85	3.85
	Timed	-	47	3.50
	Untimed	-	38	4.27
	-	4	28	3.03
	-	8	57	4.25
	Timed	4	16	2.77
	Timed	8	31	3.88
	Untimed	4	12	3.38
	Untimed	8	26	4.68

Table 4
Pretest Coefficient α Reliabilities*

	Test Level	# of Items	Pretest α for Timed Group*	Pretest α for Untimed Group*	Pretest α for Timed and Untimed Groups Combined
Grade 4	9	36	.46 (N = 15)	-.50 (N = 11)	.36 (N = 26)
Grade 8	13	46	.17 (N = 9)	.00 (N = 12)	.00 (N = 21)

**All* students in this study took the pretest without time limits.

Table 5
Posttest Coefficient α Reliabilities

	Test Level	# of Items	Posttest α for Timed Group*	Posttest α for Untimed Group*	Posttest α for Timed and Untimed Groups Combined
Grade 4	9	44	.26 (N = 15)	.64 (N = 11)	.41 (N = 26)
Grade 8	13	57	.17 (N = 9)	.70 (N = 12)	.62 (N = 21)